

Neural Discourse Relation Recognition with Semantic Memory

Biao Zhang^{1,2}, Deyi Xiong² and Jinsong Su¹

Xiamen University, Xiamen, China 361005¹

Soochow University, Suzhou, China 215006²

zb@stu.xmu.edu.cn, jssu@xmu.edu.cn

dyxiong@suda.edu.cn

Abstract

Humans comprehend the meanings and relations of discourses heavily relying on their semantic memory that encodes general knowledge about concepts and facts. Inspired by this, we propose a neural recognizer for implicit discourse relation analysis, which builds upon a semantic memory that stores knowledge in a distributed fashion. We refer to this recognizer as *SeMDER*. Starting from word embeddings of discourse arguments, *SeMDER* employs a *shallow encoder* to generate a distributed surface representation for a discourse. A *semantic encoder* with attention to the semantic memory matrix is further established over surface representations. It is able to retrieve a deep semantic meaning representation for the discourse from the memory. Using the surface and semantic representations as input, *SeMDER* finally predicts implicit discourse relations via a *neural recognizer*. Experiments on the benchmark data set show that *SeMDER* benefits from the semantic memory and achieves substantial improvements of 2.56% on average over current state-of-the-art baselines in terms of F1-score.

1 Introduction

Discourse relation recognition (DRR) that automatically identifies the logical relation of a coherent text is very important for discourse-level comprehension. It is relevant to a variety of nature language processing tasks such as summarization [Yoshida *et al.*, 2014], machine translation [Guzmán *et al.*, 2014], question answering [Jansen *et al.*, 2014] and information extraction [Cimiano *et al.*, 2005]. Although explicit DRR has recently achieved remarkable success [Mitsakaki *et al.*, 2005; Pitler *et al.*, 2008], implicit DRR still remains a serious challenge due to the absence of discourse connectives.

However, even if discourse connectives are not provided, humans can still easily succeed in recognizing the relations of discourse arguments. One reason for this, according to cognitive psychology, would be that humans have a semantic memory in mind, which helps them comprehend word senses and further argument meanings via composition. After understanding what two arguments of a discourse convey, humans can easily interpret the discourse relation of the two

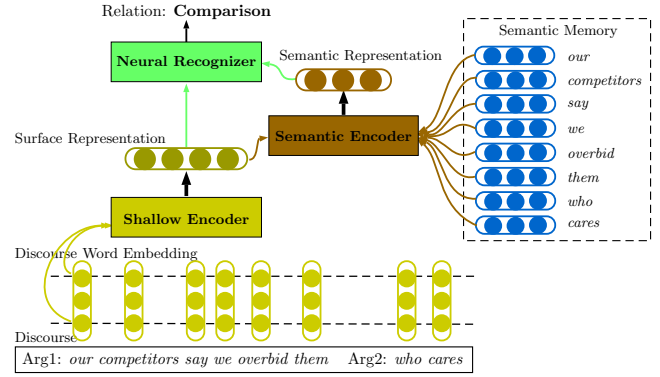


Figure 1: Overall architecture for *SeMDER* model. We use the shallow and deep yellow color to indicate the surface and semantic representation respectively.

arguments. This semantic memory, as discussed by Tulving [1972], refers to general knowledge including “words and other verbal symbols, their meaning and referents, about relations among them, and about rules, formulas, and algorithms for manipulating them”. It can be retrieved to help disambiguation and comprehension whenever the barrier of cognition occurs.

Consider the implicit discourse relation between the following two sentences:

- (1) *I was prepared to be in a very bad mood tonight.*
Now, I feel maybe there’s a little bit of euphoria.

It is difficult for conventional discourse relation recognizers to identify the relation between the two sentences as there is little significant surface information for use. However, if the recognizer obtains the knowledge of the antonymous relationship between the meaning of “*bad mood*” and that of “*euphoria*”, it will be easy to infer the COMPARISON relation between the two sentences. This semantic knowledge can be stored in an external memory for a discourse recognizer just like the semantic memory for humans.

Inspired by the semantic memory in cognitive neuroscience [Yee *et al.*, 2014] as well as memory network [Weston *et al.*, 2014; Sukhbaatar *et al.*, 2015; Kumar *et al.*, 2015] and attentional mechanisms [Mnih *et al.*, 2014; Bahdanau *et al.*, 2014; Xu *et al.*, 2015], we propose a neural network with se-

semantic memory for implicit DRR, which refers to SeMDER. The philosophy of SeMDER includes: (1) the external semantic memory should be distributed as this allows easy computation; (2) the semantic memory should be easily accessed and retrieved; and (3) the retrieved content should be integrated into the comprehension of meanings of discourse arguments and their relations. In order to meet these requirements, we use a distributed matrix that encodes semantic knowledge of words as our external memory. The distributed memory is retrieved via an attentive reader. The retrieved distributed knowledge is incorporated into semantic representations of discourse arguments. Practically, we build a neural network that is composed of three essential components: a shallow encoder, a semantic encoder and a neural recognizer. The neural network is visualized in Figure 1. In particular,

- **Shallow encoder:** we feed word embeddings of discourse arguments into a *shallow encoder* [Zhang *et al.*, 2015] to obtain shallow representations of arguments. Due to their shallow property, we refer to them as surface representations (see Section 3.1);
- **Semantic encoder:** we retrieve the semantic memory via an attention model. The retrieved content, together with surface representations, are incorporated into the semantic encoder to obtain deep semantic representations (see Section 3.2);
- **Neural recognizer:** both surface and semantic representations are feed into a *neural recognizer* to predict the corresponding discourse relations (see Section 3.3).

Our contributions are twofold. First, we propose a neural network architecture for implicit DRR with an encoded semantic memory that enhances representations of arguments. To the best of our knowledge, we are the first to explore semantic memory for DRR via attentional mechanisms. Second, we conduct a series of experiments for English implicit DRR on the PDTB-style corpus to evaluate the effectiveness of our proposed neural network and semantic memory. Experiment results show that our network achieves substantial improvements against several strong baselines in term of F1 score. Extensive analysis on the attention further indicates that our model can recognize some important relation-relevant words, which we conjecture is the main reason for our success.

2 Related Work

The release of Penn Discourse Treebank (PDTB) [Prasad *et al.*, 2008] opens the door to machine learning based implicit DRR. A variety of machine learning strategies have been presented previously, including feature engineering, connective predicting, data selection and discourse representation via neural networks.

Research on feature engineering exploits powerful and discriminative features for implicit DRR. In this respect, Pilter *et al.* [2009] investigate several linguistically informed features, such as polarity tags, verb classes, modality, context and lexical features. Lin *et al.* [2009] further consider contextual words, word pairs and parse trees for feature engineering. Later, several more powerful features have been developed:

aggregated word pairs [McKeown and Biran, 2013], Brown clusters and coreference patterns [Rutherford and Xue, 2014]. With these features, Park and Cardie [2012] perform feature set optimization for better feature combination.

The major difference between explicit and implicit DRR is the presence of discourse connectives, the most salient features for DRR. Therefore, if we find a way to predict connectives for implicit discourses, we can transform implicit DRR into explicit DRR. Along this line, Zhou *et al.* [2010] use a language model to automatically insert discourse connectives, while Patterson and Kehler [2013] use a classifier to predict the presence or omission of a lexical connective. Different from this prediction strategy, Hong *et al.* [2012] leverage discourse connectives as a bridge between explicit and implicit relations and adopt an unsupervised cross-argument inference mechanism.

Yet another strategy is data selection, where explicit discourse instances that are similar to the implicit ones are found and added to training corpus. Different data selection methods for implicit DRR can be classified into the following categories: instance typicality [Wang *et al.*, 2012], multi-task learning [Lan *et al.*, 2013], domain adaptation [Braud and Denis, 2014; Ji *et al.*, 2015], semi-supervised learning [Hernault *et al.*, 2010; Fisher and Simmons, 2015] and explicit discourse connective classification [Rutherford and Xue, 2015].

The third strategy is to learn representations of discourse arguments using neural networks for relation recognition, following remarkable success of neural networks in various natural language processing tasks. In this respect, Braud and Denis [2015] investigates the usefulness of word representations. Specifically, two different neural network models have been developed for implicit DRR: recursive neural network for entity-augmented distributed semantics [Ji and Eisenstein, 2015] and shallow convolutional neural network for discourse representation [Zhang *et al.*, 2015]. The former incorporates coreferent entity mentions into compositional distributed representations, while the latter develops a pure neural network model for discourse representations in implicit DRR. Normally, entities utilized in the former heavily depend on the availability and robustness of an upstream coreference system, and the latter only learns shallow representations for discourse arguments. Instead, our proposed model does not rely on any linguistic resources and incorporates a semantic memory to obtain deep semantic representations over shallow representations in [Zhang *et al.*, 2015]. Additionally, since the semantic memory is represented as a distributed matrix, our model is more robust and adaptable.

The exploration of semantic memory for implicit DRR is inspired by recent developments in cognitive neuroscience. Yee *et al.* [2014] show how this memory is organized and retrieved in brain. In order to explore semantic memory in neural networks, we borrow ideas from recently introduced memory networks [Weston *et al.*, 2014; Sukhbaatar *et al.*, 2015; Kumar *et al.*, 2015] to organize semantic memory as a distributed matrix and use an attention model to retrieve this distributed memory. The adaptation and utilization of semantic memory into implicit DRR, to the best of our knowledge, has never been investigated before.

3 The SeMDER Model

This section elaborates the proposed SeMDER model. We will first present the shallow encoder which converts a discourse into a distributed embedding. The semantic encoder, where the semantic memory is incorporated via an attention model is then described. After that, we explain how the neural recognizer classifies discourse relations. We also discuss the objective function and the procedure for parameter learning in this section.

3.1 Shallow Encoder

To obtain surface representations for discourses, we employ a shallow convolutional neural network (SCNN) [Zhang *et al.*, 2015] as our shallow encoder. SCNN is specifically designed on the PDTB corpus, where implicit discourse relations are annotated between two neighboring arguments, namely *Arg1* and *Arg2* (see the example in Figure 1). Given an argument which consists of n words, SCNN represents each word as a d -dimensional dense, real-valued vector $x_i \in \mathbb{R}^{d_1}$ and concatenates them into a word embedding matrix:

$$X = (x_1, x_2, \dots, x_n) \quad (1)$$

where $X \in \mathbb{R}^{d_1 \times n}$ forms the input layer of SCNN. All word vectors in vocabulary V are stacked into a parameter matrix $L \in \mathbb{R}^{d_1 \times |V|}$ ($|V|$ is the vocabulary size), which will be tuned in the training phase.

To represent a discourse argument c , SCNN extracts major information inside X through three convolution operations *avg*, *min* and *max* defined as follows:

$$c_r^{avg} = \frac{1}{n} \sum_i^n X_{r,i} \quad (2)$$

$$c_r^{min} = \min(X_{r,1}, X_{r,2}, \dots, X_{r,n}) \quad (3)$$

$$c_r^{max} = \max(X_{r,1}, X_{r,2}, \dots, X_{r,n}) \quad (4)$$

where r indicates the row of X . The argument c is thereby represented as the concatenation of these convolutional features:

$$c = [c^{avg}; c^{max}; c^{min}] \quad (5)$$

SCNN further obtains the representation $p \in \mathbb{R}^{6d_1}$ of a discourse by performing nonlinear transformations on the concatenation of two argument embeddings c_{Arg1} and c_{Arg2} generated in Eq. 5 as follows:

$$p = g([c_{Arg1}; c_{Arg2}]), \quad g(x) = \frac{\tanh(x)}{\|\tanh(x)\|} \quad (6)$$

Despite its simplicity, SCNN outperforms several feature-based systems. This is the reason that we choose it as our shallow encoder to obtain surface representations of discourses. However, the lack of deep knowledge in SCNN limits its further development. We therefore introduce a deep semantic encoder over the shallow encoder, which will be elaborated in the next section.

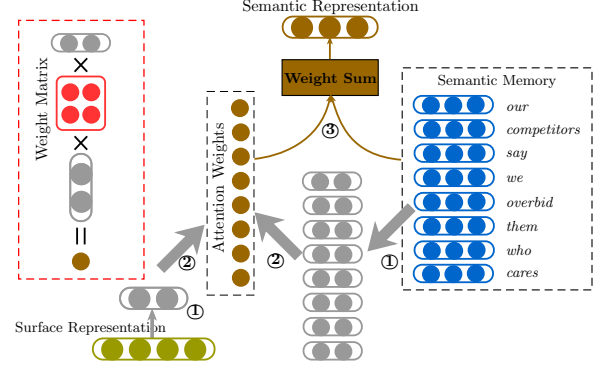


Figure 2: Illustration of the semantic encoder. We use gray color to indicate representations in the attention space. The dashed red box shows the bilinear-style computation for attention weights.

3.2 Semantic Encoder

Upon the surface representations, we further build a semantic encoder to incorporate a semantic memory to strengthen discourse comprehension. The semantic memory in SeMDER is represented as a distributed matrix $M \in \mathbb{R}^{m \times d_2}$, where d_2 is the dimension of word embedding in the memory. Each row in the matrix indicates one word in discourse arguments (thus typically $m \leq n$). We assume that the semantic and syntactic attributes of words have already been encoded into this matrix. Therefore, incorporating this memory information into discourse representations will be beneficial for implicit DRR task.

Figure 2 gives an illustration of the procedure for incorporating the semantic memory. Specifically, given the surface representation p for a discourse and the semantic memory matrix M , we stack an attention layer to project them onto the same space, which we call attention space. The projection is done as follows:

$$p_a = f(W_p p + b_a) \quad (7)$$

$$M_a = f(W_m M^T + b_a) \quad (8)$$

where the subscript a denotes the attention space, p_a and M_a are the attentional representations for p and M respectively. $W_p \in \mathbb{R}^{d_a \times 6d_1}$, $W_m \in \mathbb{R}^{d_a \times d_2}$ are transformation matrices, $b_a \in \mathbb{R}^{d_a}$ is the bias term, d_a is the dimensionality of the attention space, and $f(\cdot)$ is an element-wise activation function such as $\tanh(\cdot)$, which is used throughout our model. The arrows marked by "①" in Figure 2 show this projection process.

Note that we differentiate the transformation matrix W_p in Eq. 7 to the W_m in Eq. 8, since the surface representation and semantic memory are from different semantic spaces. However, we share the same bias term for them. This will force our model to learn to encode attention semantics into the transformation matrices, rather than the biases.

After obtaining the attentional representations for the discourse and semantic memory, we further estimate how useful each word memory cell i in the semantic memory (i.e., the i th row in M) is to the corresponding discourse. This can be

calculated by a match score:

$$s_i = g(p_a, M_{a,i}) \quad (9)$$

where $g(\cdot)$ is the scoring function. Since we are only interested in words occurred in the corresponding discourse, our attention schema is somewhat like a local attention. As discussed in [Luong *et al.*, 2015], a *general* scoring function is much better for the local attention. Thus, we use a variant of the *general* function as our scoring function (see the red box in Figure 2):

$$g(p_a, M_{a,i}) = p_a W_s M_{a,i} \quad (10)$$

where $W_s \in \mathbb{R}^{d_a \times d_a}$ is the bilinear scoring matrix, in which each element (see the red node in Figure 2) represents an interaction between the corresponding dimension of p_a and $M_{a,i}$.

We further normalize the match score vector in Eq. 9 to generate a probabilistic attention distribution over words in the semantic memory:

$$\alpha_i = \frac{\exp(s_i)}{\sum_{j=1}^m \exp(s_j)} \quad (11)$$

Intuitively, the probability α_i (a.k.a attention weight) reflects the importance of the word M_i in representing the whole discourse with respect to the final discourse relation recognition. Recall the above-mentioned example (1), if the importance of words “*bad mood*” and “*euphoria*” is recognized, there would be more chance that the final recognizer succeeds.

Based on this attention distribution, we can compute the semantic representation for a discourse as a weighted sum of words in the semantic memory according to α (see the arrows marked by “③” in Figure 2):

$$p_k = \sum_{j=1}^m \alpha_j M_j \quad (12)$$

As shown in Eq. 12, the semantic representation is directly retrieved from the semantic memory. It encodes semantic knowledge of words in discourse arguments that can help discourse relation recognition.

3.3 Neural Recognizer

Up to now, we have inferred both the surface and semantic representation for a discourse. To recognize the discourse relation, we further stack a Softmax layer upon these two representations:

$$y_p = h(W_{r,p}p + W_{r,k}p_k + b_r) \quad (13)$$

where $h(\cdot)$ is the softmax function, $W_{r,p} \in \mathbb{R}^{l \times 6d_1}$, $W_{r,k} \in \mathbb{R}^{l \times d_2}$ and $b_r \in \mathbb{R}^l$ are the parameter matrices and bias term respectively, and l indicates the number of discourse relations.

3.4 Objective Function and Parameter Learning

Given a training corpus which contains T instances $\{(x, y)\}_{t=1}^T$, we employ the following cross-entropy error to

access how well the predicted relation y_p represents the gold relation y ,

$$E(y_p, y) = - \sum_j^l y_j \times \log(y_{p,j}) \quad (14)$$

Therefore, the joint training objective of SeMDER is defined as follows:

$$J(\theta) = \frac{1}{T} \sum_{t=1}^T E(y_p^{(t)}, y^{(t)}) + R(\theta) \quad (15)$$

where $R(\theta)$ is the regularization term with respect to θ . Towards the parameters θ , we divide them into three different sets:

- θ_L : word embedding matrix L ;
- θ_R : discourse relation recognition parameters $W_{r,p}$, $W_{r,k}$ and b_r ;
- θ_M : memory-related parameters W_p , W_m , W_s and b_a ;

All these parameters are regularized with corresponding weights¹:

$$R(\theta) = \frac{\lambda_L}{2} \|\theta_L\|^2 + \frac{\lambda_R}{2} \|\theta_R\|^2 + \frac{\lambda_M}{2} \|\theta_M\|^2 \quad (16)$$

Notice that although we can fine-tune the semantic memory in an end-to-end manner, we do not do that in our model. This is because we hope that the semantic and syntactic attributes encoded in the semantic memory can be preserved throughout our neural network.

We apply Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm to optimize each parameter. In order to run the L-BFGS algorithm, we need to solve two problems: parameter initialization and partial gradient calculation.

In the phase of parameter initialization, θ_R and θ_M are randomly set according to a normal distribution ($\mu = 0, \sigma = 0.01$). For the word embedding θ_L , we use the toolkit Word2Vec² to perform pretraining on a large-scale unlabeled data. This word embedding will be further fine-tuned in our SeMDER model to capture much more semantics related to discourse relations.

The partial gradient for parameter θ_j is computed as follows:

$$\frac{\partial J}{\partial \theta_j} = \frac{1}{T} \sum_{t=1}^T \frac{\partial E(y_p^{(t)}, y^{(t)})}{\partial \theta_j} + \lambda_j \theta_j \quad (17)$$

This gradient will be feed into the toolkit libLBFGS³ for parameter updating in our practical implementation.

4 Experiments

In this section, we conducted a series of experiments on English implicit DRR task. We begin with a brief review of the PDTB dataset. Then, we describe our experiment setup. Finally, we present experiment results and give an in-depth analysis on the attention.

¹The bias terms b is not regularized in practice.

²<https://code.google.com/p/word2vec/>

³<http://www.chokkan.org/software/liblbfgs/>

Relation	Positive/Negative Sentences		
	Train	Dev	Test
COM	1942/1942	197/986	152/894
CON	3342/3342	295/888	279/767
EXP	7004/7004	671/512	574/472
TEM	760/760	64/1119	85/961

Table 1: Statistics of implicit discourse relations for the training (Train), development (Dev) and test (Test) sets in PDTB corpus.

4.1 Dataset

We used *PDTB 2.0* corpus⁴ [Prasad *et al.*, 2008] (PDTB thereafter), which is the largest hand-annotated discourse corpus. Discourse relations are annotated in a predicate-argument view in PDTB, where each discourse connective is treated as a predicate that takes two text spans as its arguments. The relation tags in PDTB are arranged in a three-level hierarchy, where the top level consists of four major semantic classes: TEMPORAL (TEM), CONTINGENCY (CON), EXPANSION (EXP) and COMPARISON (COM). Because the top-level relations are general enough to be annotated with a high inter-annotator agreement and are common to most theories of discourse, in our experiments we only use this level of annotations.

PDTB contains discourse annotations over 2,312 Wall Street Journal articles, and is organized in different sections. Following previous work [Pitler *et al.*, 2009; Zhou *et al.*, 2010; Lan *et al.*, 2013; Zhang *et al.*, 2015], we used sections 2-20 as our training set, sections 21-22 as the test set. Sections 0-1 were used as the development set for hyperparameter optimization. We formulated the task as four separate one-against-all binary classification problems: each top level class vs. the other three discourse relation classes. We also balanced the training set by resampling training instances in each class until the number of positive and negative instances are equal. In contrast, all instances in the test and development set are kept in nature. The statistics of various data sets is listed in Table 1.

4.2 Setup

We selected the *GoogleNews-vectors-negative300*⁵ as our external semantic memory. This data contains 300-dimensional vectors (thus, $d_2 = 300$) for 3 million words and phrases. It is trained on part of Google News dataset (about 100 billion words). The wide coverage and newswire domain of its training corpus as well as the syntactic property of word2vec models make this vector a good choice for the semantic memory.

We tokenized all datasets using *Stanford NLP Toolkit*⁶, and employed a large-scale unlabeled data⁷ including 1.02M

⁴<http://www.seas.upenn.edu/pdtb/>

⁵<https://drive.google.com/file/d/0B7XkCwpI5KDYNNUTTISS21pQmM/edit?pref=2&pli=1>

⁶<http://nlp.stanford.edu/software/corenlp.shtml>

⁷This data contains the training and development set for implicit DRR, as well as the English sentences in the FBIS corpus and the English sentences in Hansards part of LDC2004T07 corpus.

sentences (33.5M words) for word embedding θ_L initialization. We optimized the hyperparameters $d_1, \lambda_L, \lambda_R, \lambda_M$ according to previous work [Zhang *et al.*, 2015] and preliminary experiments on the development set. Finally, we set $d_1 = 128, \lambda_L = 1e^{-5}, \lambda_R = \lambda_M = 1e^{-4}$ for all the experiments. With respect to d_a , we tried three different settings $d_a = 32, 64, 128$.

To validate the effectiveness of **SeMDER** model, we compared against the following baseline methods:

- **SVM**: a support vector machine (SVM) classifier trained with the labeled data in the training set. We used the toolkit *SVM-light*⁸ to train the classifier in our experiments.
- **SCNN**: a shallow convolutional neural model proposed by Zhang *et al.* [2015].

Features used in **SVM** experiments are taken from the state-of-the-art implicit discourse relation recognition model, including *Bag of Words*, *Cross-Argument Word Pairs*, *Polarity*, *First-Last*, *First3*, *Production Rules*, *Dependency Rules* and *Brown cluster pair* [Rutherford and Xue, 2014]. Additionally, in order to collect bag of words, production rules, dependency rules, and cross-argument word pairs, we used a frequency cutoff of 5 to remove rare features, following Lin *et al.* [2009].

4.3 Classification Results

Because of the imbalance nature in our test set, we choose F1 score as our major evaluation metric. The performance of different models is presented in Table 2, which, overall, shows that **SeMDER** outperforms the two baselines, achieving improvements in F1 score of 1.14% on COM, 1.66% on CON, 1.36% on EXP and 5.62% on TEM over the best baseline results. We further observe that the improvements mainly result from high precision for COM, CON and TEM, while high recall for EXP. This is reasonable since the EXP relation owns the largest number of instances in our data.

As the neural baseline, **SCNN** outperforms **SVM** on CON, EXP and TEM, but fails on COM. The **SeMDER** with semantic memory, however, consistently surpasses **SVM** and **SCNN** in all discourse relations. This suggests that the incorporated semantic memory is helpful for recognizing correct discourse relations. Additionally, for **SeMDER**, increasing the attention space dimensionality d_a from 32 to 128 improves the performance in most cases.

Yet another interesting observation from Table 2 is that the improvement of **SeMDER** over the two baselines for relation TEM is the biggest. The gain over **SVM** is 11.4% and 5.6% over **SCNN**. This improvement is largely due to high precisions. As the number of instances in relation TEM is the smallest (see Table 1), we argue that the traditional neural network models may suffer from overfitting in this case. However, our **SeMDER** enhanced with the semantic memory is capable of generalization that alleviates this overfitting issue.

⁸<http://svmlight.joachims.org/>

Model	d_a	P	R	F1
SVM	-	22.79	64.47	33.68
SCNN	-	22.00	67.76	33.22
SeMDER	32	22.18	73.68	34.09
	64	23.33	61.84	33.87
	128	25.71	53.95	34.82

(a) COM vs Other

Model	d_a	P	R	F1
SVM	-	65.89	58.89	62.19
SCNN	-	56.29	91.11	69.59
SeMDER	32	54.80	99.48	70.67
	64	54.79	99.65	70.70
	128	54.98	100.0	70.95

(c) EXP vs Other

Model	d_a	P	R	F1
SVM	-	39.14	72.40	50.82
SCNN	-	39.80	75.29	52.04
SeMDER	32	41.14	74.91	53.11
	64	39.82	80.65	53.32
	128	42.07	74.19	53.70

(b) CON vs Other

Model	d_a	P	R	F1
SVM	-	15.10	68.24	24.73
SCNN	-	20.22	62.35	30.54
SeMDER	32	21.79	60.00	31.97
	64	23.01	61.18	33.44
	128	34.78	37.65	36.16

(d) TEM vs Other

Table 2: Classification results of different models on implicit DRR. **P**=Precision, **R**=Recall, and **F1**=F1 score. The best F1 scores are highlighted in bold.

Relation	Example	Top Words
COM	[people think of the steel business as an old and mundane smokestack business] _{Arg1} , [they 're dead wrong] _{Arg2}	<i>wrong, people, dead, think, smokestack</i>
CON	[three minutes into the massage , the man curled up , began shaking and turned red] _{Arg1} , [paramedics were called] _{Arg2}	<i>shaking, turned, paramedics, massage, curled</i>
EXP	[numerous injuries were reported] _{Arg1} , [some buildings collapsed , gas and water lines ruptured and fires raged] _{Arg2}	<i>injuries, were, collapsed, raged, ruptured</i>
TEM	[warner sued sony and guber-peters late last week] _{Arg1} , [sony and guber-peters have countersued] _{Arg2}	<i>have, countersued, late, week, last</i>

Table 3: Attention examples selected from the test set (we set $d_a = 128$ for all relations). The top words are arranged in the order of attention weights.

4.4 Attention Analysis

We would like to know more about what role the semantic memory plays in our model, especially what the model learns from this semantic memory. Analyzing semantic representations is relatively meaningless. Therefore we turn to look into words with high attention weights for the answer.

We present one example per discourse relation from the test set in Table 3, where words assigned with the top-5 attention weights are listed separately. Consider the example for COM, our model retrieves the words “*wrong, people, dead, think, smokestack*”, which roughly reflect the discourse meaning that *people think smokestack, dead wrong*. Obviously, these words are crucial for discourse comprehension. These examples display that SeMDER prefers to retrieve from the semantic memory relation-relevant words that strongly indicate the corresponding relations, which we think is the main reason for the success of SeMDER.

5 Conclusion and Future Work

In this paper, we have presented a neural discourse relation recognizer with a distributed semantic memory for implicit DRR. The semantic memory encodes semantic knowledge of words in discourse arguments and helps disambiguation and comprehension. We employ an attention model to retrieve

discourse relation-relevant information into semantic representations of discourses, which, to some extend, simulates the cognition process of humans. Experiment results show that our model outperforms several strong baselines, and further analysis reveals that our model can indeed detect some relation-relevant words.

In the future, we would like to exploit different types of semantic memory, e.g., a distributed memory on ontology concepts and relations. We also want to explore different attention architectures, e.g. the *concat* and *dot* in [Luong *et al.*, 2015]. Furthermore, we are interested in adapting our model to other similar classification tasks, such as sentiment classification, movie review classification and nature language inference.

References

- [Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. 2014.
- [Braud and Denis, 2014] Chloé Braud and Pascal Denis. Combining natural and artificial examples to improve implicit discourse relation identification. In *Proc. of COLING*, pages 1694–1705, August 2014.
- [Cimiano *et al.*, 2005] Philipp Cimiano, Uwe Reyle, and Jasmin Šarić. Ontology-driven discourse analysis for information extraction. *Data & Knowledge Engineering*, 55:59–83, 2005.
- [Fisher and Simmons, 2015] Robert Fisher and Reid Simmons. Spectral semi-supervised discourse relation classification. In *Proc. of ACL-IJCNLP*, pages 89–93, July 2015.
- [Guzmán *et al.*, 2014] Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. Using discourse structure improves machine translation evaluation. In *Proc. of ACL*, pages 687–698, June 2014.
- [Hernault *et al.*, 2010] Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka. A Semi-Supervised Approach to Improve Classification of Infrequent Discourse Relations Using Feature Vector Extension. In *Proc. of EMNLP*, 2010.
- [Hong *et al.*, 2012] Yu Hong, Xiaopei Zhou, Tingting Che, Jianmin Yao, Qiaoming Zhu, and Guodong Zhou. Cross-argument inference for implicit discourse relation recognition. In *Proc. of CIKM*, pages 295–304, 2012.
- [Jansen *et al.*, 2014] Peter Jansen, Mihai Surdeanu, and Peter Clark. Discourse complements lexical semantics for non-factoid answer reranking. In *Proc. of ACL*, pages 977–986, June 2014.
- [Ji and Eisenstein, 2015] Yangfeng Ji and Jacob Eisenstein. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *TACL*, pages 329–344, 2015.
- [Ji *et al.*, 2015] Yangfeng Ji, Gongbo Zhang, and Jacob Eisenstein. Closing the gap: Domain adaptation from explicit to implicit discourse relations. In *Proc. of EMNLP*, pages 2219–2224, September 2015.
- [Kumar *et al.*, 2015] Ankit Kumar, Ozan Irsoy, Jonathan Su, James Bradbury, Robert English, Brian Pierce, Peter Ondruska, Ishaan Gulrajani, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. *CoRR*, 2015.
- [Lan *et al.*, 2013] Man Lan, Yu Xu, and Zhengyu Niu. Leveraging Synthetic Discourse Data via Multi-task Learning for Implicit Discourse Relation Recognition. In *Proc. of ACL*, pages 476–485, Sofia, Bulgaria, August 2013.
- [Lin *et al.*, 2009] Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proc. of EMNLP*, pages 343–351, 2009.
- [Luong *et al.*, 2015] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. *Proc. of EMNLP*, 2015.
- [McKeown and Biran, 2013] Kathleen McKeown and Or Biran. Aggregated word pair features for implicit discourse relation disambiguation. In *Proc. of ACL*, pages 69–73, 2013.
- [Miltsakaki *et al.*, 2005] Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Experiments on sense annotations and sense disambiguation of discourse connectives. In *Proc. of TLT2005*, 2005.
- [Mnih *et al.*, 2014] Volodymyr Mnih, Nicolas Heess, Alex Graves, and koray kavukcuoglu. Recurrent models of visual attention. In *Proc. of NIPS*, pages 2204–2212, 2014.
- [Park and Cardie, 2012] Joonsuk Park and Claire Cardie. Improving Implicit Discourse Relation Recognition Through Feature Set Optimization. In *Proc. of SIGDIAL*, pages 108–112, Seoul, South Korea, July 2012.
- [Patterson and Kehler, 2013] Gary Patterson and Andrew Kehler. Predicting the presence of discourse connectives. In *Proc. of EMNLP*, pages 914–923, 2013.
- [Pitler *et al.*, 2008] Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind K Joshi. Easily identifiable discourse relations. *Technical Reports (CIS)*, page 884, 2008.
- [Pitler *et al.*, 2009] Emily Pitler, Annie Louis, and Ani Nenkova. Automatic sense prediction for implicit discourse relations in text. In *Proc. of ACL-AFNLP*, pages 683–691, August 2009.
- [Prasad *et al.*, 2008] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. The penn discourse treebank 2.0. In *LREC*. Citeseer, 2008.
- [Rutherford and Xue, 2014] Attapol Rutherford and Nianwen Xue. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proc. of EACL*, pages 645–654, April 2014.
- [Rutherford and Xue, 2015] Attapol Rutherford and Nianwen Xue. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *Proc. of NAACL-HLT*, pages 799–808, May–June 2015.
- [Sukhbaatar *et al.*, 2015] Sainbayar Sukhbaatar, arthur szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In *Proc. of NIPS*, pages 2431–2439, 2015.
- [Tulving, 1972] Endel Tulving. Episodic and semantic memory. In Endel Tulving and W. Donaldson, editors, *Organization of Memory*, pages 381–403. Academic Press, New York, 1972.
- [Wang *et al.*, 2012] Xun Wang, Sujian Li, Jiwei Li, and Wenjie Li. Implicit discourse relation recognition by selecting typical training examples. In *Proc. of COLING*, pages 2757–2772, 2012.
- [Weston *et al.*, 2014] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *CoRR*, abs/1410.3916, 2014.
- [Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. of ICML*, pages 2048–2057, 2015.
- [Yee *et al.*, 2014] Eiling Yee, Evangelia G Chrysikou, and Sharon L Thompson-Schill. The cognitive neuroscience of semantic memory, 2014.
- [Yoshida *et al.*, 2014] Yasuhisa Yoshida, Jun Suzuki, Tsutomu Hirao, and Masaaki Nagata. Dependency-based discourse parser for single-document summarization. In *Proc. of EMNLP*, pages 1834–1839, October 2014.
- [Zhang *et al.*, 2015] Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. Shallow convolutional neural network for implicit discourse relation recognition. In *Proc. of EMNLP*, September 2015.
- [Zhou *et al.*, 2010] Zhi-Min Zhou, Yu Xu, Zheng-Yu Niu, Man Lan, Jian Su, and Chew Lim Tan. Predicting discourse connectives for implicit discourse relation recognition. In *Proc. of COLING*, pages 1507–1514, 2010.